Feature Extraction and Evaluation of Electricity Load Data with High Precision

Jiangqi Chen[1], Ting Zhao[1], Yang Yang[2], Di Zhang[1]
[1]Advanced Computing and Big Data Technology Laboratory of SGCC, Global Energy Interconnection Research Institute, Beijing, China
[2]School of Control and Computer Engineering, North China Electric Power University, Beijing, China
e-mail: chenjiangqi@geiri.sgcc.com.cn

***Abstract*—This paper summarizes the characteristics of the electricity load data collected every 15 minutes of power users. The single-day load data of single power user is taken as a 96-length vector, and the single-day data of n users is taken as a 96-column set. *Single*-user monthly data and annual data are described as a 30 (31) × 96 matrix and a 365 (366) × 96 matrix respectively. By comparing the wavelet transforms of multiple wavelet basis functions, the Daubechies 4 wavelet basis function is chosen as the wavelet transform basis function of the electricity load data according to the Shannon entropy of the obtained wavelet coefficients. The difference between the inverse wavelet transformed data and the original data is compared under various characteristic wavelet coefficients. Besides, the number of the characteristic wavelet coefficients and the error of the wavelet transform are determined. The results show that there is a high redundancy in the original power load data, and the storage space can be greatly compressed by wavelet transform to achieve feature extraction and data desensitization. Therefore, the feature extraction technology of power load data based on wavelet transform in this paper has potential application value and popularization value, and can produce high economic efficiency.**

Keywords-wavelet transform; wavelet multi-resolution decomposition; electric information acquisition data

I. INTRODUCTION

At present, the load data of power users has become an important data source for government to carry out smart city construction and for telecom operators, Internet companies, finance and insurance industry to make customer portraits, as well as for retail entertainment enterprises to assist commercial location. However, due to the large number of meters, the high sampling frequency and the huge data amount, a network with large storage capacity, high bandwidth and high computing performance is in great need to perform data handle, storage, process and analysis. Therefore, a high precision feature extraction of load data in application of data mining algorithm is actually in urgent need.

At present, the methods of data characterization are mainly based on three categories: statistics, model and transformation. Among which, the characterization based on transformation refers to the highlight of classification characteristic by means of transformation, including time-frequency transformation and linear transformation [1-3]. Representative time-frequency conversion methods include fast Fourier transform, short-time Fourier transform and cepstral coefficients. The classic linear transformation is principal component analysis (PCA), singular value decomposition (SVD), linear discriminant analysis, K-L transform, wavelet transform and wavelet packet technology [4-6].

The generalized empirical wavelet transform is used to reproduce the single and composite power quality distribution, which has a good effect on the precise classification of power quality [7]. The time series is transformed from time domain to frequency domain through signal processing, and the finite time series in the frequency domain is employed to approximately describe the original time series [8]. In paper [9], an approximate optimal histogram construction algorithm based on workload is proposed, a method of calculating the unselected wavelet coefficients under the L_2 error metric and min-max error metric is presented, and an improved table semantic aggregation method and that of graph is studied.

However, whether the wavelet transform is suitable for the feature extraction of the electricity load data, whether can it ensure the characteristic can indicate users' power consumption trend, and whether the detailed electricity information can be expressed by means of flexible selection of the characteristic coefficients are all under consideration.

In general, this paper studies the structure feature of power users' electricity consumption behavior, and selects the appropriate wavelet basis function to carry out multi-resolution decomposition of electricity load data. Thus the electrical information characteristics of each power user are extracted to achieve data compression and data desensitization. In addition, this paper analyzes the loss of the original data caused by wavelet transform through the inverse wavelet transform of the characteristic value of electricity information, thus the feasibility of the feature extraction method studied above is evaluated.

II. EXTRACTION AND SELECTION OF WAVELET FEATURE

A. Wavelet Transform

In application of Wavelet Transform (WT) to analyze the original signal, the dual localization property in both time domain and frequency domain could be achieved. The corresponding sampling step is applied for different signals in time domain, so as to focus on any slight detail of the signal. WT is considered to achieve better effect than Fourier Transform and windowed Fourier Transform, and a

breakthrough of modern Fourier analysis. Besides, it is praised as "the mathematical microscope" [10-13].

As with a square integrable function $\psi(t)$, if $\int_{-\infty}^{+\infty} \psi(t)\, dt = 0$, then $\psi(t)$ is called a wavelet. Typically a set of wavelets scaled and translated from single wavelet function is applied during a wavelet transform. As defined by

$$\psi_{a,b}(t) = |a|^{\frac{1}{2}} \psi\left(\frac{t-b}{a}\right), \ (a, b \in R, a \neq 0) \tag{1}$$

where a is the scale parameter, b is the translation parameter and ψ is called the basic wavelet or the mother wavelet.

Wavelet transform can be divided into two types, Continue Wavelet Transform (CWT) and Discrete Wavelet Transform (DWT). CWT is commonly employed for theoretical research and DWT whose scale variables and translated variables both have been discretized is more frequently used in practical application.

Define $a = a_0^j$, $b = kb_0a_0^j$, $a_0>1$, $b_0>0$, where the values of a_0 and b_0 depend on the specific form of the mother wavelet, j and k are integers. Thus, the discrete wavelet is defined as:

$$\psi_{j,k}(t) = \frac{1}{\sqrt{a_0^j}} \psi\left(\frac{t - kb_0a_0^j}{a_0^j}\right) = a_0^{-j/2} \psi(a_0^j t - kb_0) \tag{2}$$

The corresponding discrete wavelet transform is as defined by,

$$W_\psi(j,k) = \langle f(t), \psi_{j,k}(t) \rangle = a_0^{-j/2} \int_{-\infty}^{+\infty} f(t)\, \psi(a_0^{-j} t - kb_0) \mathrm{d}t \tag{3}$$

If there are positive numbers A and B, for $0<A\leqslant B< \infty$ and $\forall f(x) \in L^2(R)$, if there is

$$A\|f\|^2 \leq \sum_j \sum_k \left|\langle f, \psi_{j,k} \rangle\right|^2 \leq B\|f\|^2 \quad j, k \in Z \tag{4}$$

Then $\{\Psi_{j,k}(x)\}$ is called a wavelet framework of $L^2(R)$, And B are called frame bounds. $B<\infty$ guarantee that the transform $f \rightarrow \{\langle f, \psi_{j,k} \rangle\}$ is continuous while $A>0$ ensures that the transform is reversible and continuously inverse.

If the discrete wavelet sequence $\{\Psi_{j,k}(x)\}$ constitutes a wavelet frame, then there must be a corresponding dual sequence $\{\widetilde{\Psi}_{j,k}(x)\}$, so that $f(x)$ could simply be expressed as

$$f(x) = \sum_{j,k} \langle f(x), \Psi_{j,k}(x) \rangle \widetilde{\Psi}_{j,k}(x) = \sum_j \sum_k WT_f(j,k) \widetilde{\Psi}_{j,k}(x) \tag{5}$$

B. Multi-Resolution Decomposition

The resolution is low under large WT scale, which is suitable for general analysis while that is high under small scale, suitable for detailed observation. However, the quality factor (the ratio of the center frequency and the bandwidth) analyzed at different scales remain unchanged. The step-by-step analysis of signals is called multi-resolution decomposition analysis [14].

Let V_j denotes all the function spaces that is constant at $1/2\,j$ interval, which are all vector spaces. The scale function $\{\varphi_{j,k}; k = 0, 1, \ldots, 2^j\text{-}1\}$ is employed to constitute a base of V_j , and the satisfied nest relation is as shown

$$V_0 \subset V_1 \subset \ldots \subset V_j \subset V_{j+1} \subset \ldots \tag{6}$$

We define a function of W_j shown as $W_j=\{h \in V_{j+1}: \langle h, f \rangle=0, \forall f \in V_j\}$ as orthogonal complement of V_j in V_{j+1}, which means $V_{j+1}=V_j \oplus W_j$, in which W_j is the lost detailed information when replacing V_{j+1} with V_j, where the former indicates a function space at high resolution and the latter indicates that at low resolution. Thus, for any non-negative integer $n > 0$, the space decomposition of V_{n+1} is shown as

$$\begin{aligned} V_{n+1} &= V_n \oplus W_n \\ &= V_{n-1} \oplus W_{n-1} \oplus W_n = \ldots = V_0 \oplus W_0 \oplus W_1 \oplus \ldots \oplus W_n \end{aligned} \tag{7}$$

The multi-resolution analysis of any discrete sequence $f_{n+1} \in V_{n+1}$ is as modeled

$$f_{n+1} = f_n + d_n = f_{n-1} + d_{n-1} + d_n = \ldots = f_0 + d_0 + d_1 + \ldots + d_n \tag{8}$$

where

$f_j = \sum_{k=0}^{2^j-1} v_{j,k} \varphi_{j,k} \in V_j, j=0, 1, \ldots, n+1$,

$d_j = \sum_{k=0}^{2^j-1} w_{j,k} \psi_{j,k} \in W_j$, $j=0, 1, \ldots, n+1$, $v_{j,k}$ is the scale coefficient and $w_{j,k}$ is the wavelet coefficient respectively.

C. Selection of Wavelet Basis Function

The selection of wavelet basis function is of vital importance during data process based on WT. If the basis function and the original signal have a certain similarity, a better conversion effect could be achieved. The minimized concave cost function is basically employed to pick the best base from the base dictionary. In this way, the tools in the classical harmonic analysis like gate valve function and Shannon entropy could be used to measure the similarity between the base function and the original signal.

In this paper, the optimal wavelet basis function of electricity load data is selected based on the Shannon entropy minimum principle of wavelet decomposition coefficient. During the decomposition, if the coefficient contains the smallest entropy, then the selected basis function is the optimal wavelet basis function. The wavelet coefficients are expressed as $\{\omega_i\}$, and Shannon entropy is used to select the cost function of wavelet basis function, as defined by

$$H(\omega) = -\sum_i |\omega_i|^2 \log_2 |\omega_i|^2 \tag{9}$$

D. Extraction of Feature Wavelet Coefficients

1) Standardization of wavelet coefficients

Data characterization devotes to the simplification of data collection and transmission, data compression and data desensitization, which maximizes the intrinsic value of the

data. As a result, during feature extraction of electricity information based on WT, the data requirement of both internal business scene and external scene should be satisfied. Thus not only can useful information for business analysis be maximally kept, but also the storage space of the original data could be compressed to the most.

In this paper, to confirm the feature wavelet coefficients, initially we can sort all the wavelet coefficients according to their absolute values, then determine an optimal feature wavelet coefficient retention number B, after which the B wavelet coefficients with larger absolute values are taken as the feature coefficients of the original data, and the remaining wavelet coefficients are set to zero.

For any wavelet coefficient, the closer it is to zero, it will lose less information when set to zero. Besides, the information loss is also affected by the level of multi-resolution decomposition. For the above reasons, standardization of wavelet coefficients is essential, which is modeled as

$$Wall[i]^* = Wall[i] / \sqrt{2^{level(Wall[i])}} \tag{10}$$

where $level(x)$ indicates the level of wavelet coefficient x in multi-resolution decomposition, $Wall[i]$ indicates the i-th wavelet coefficient and $Wall[i]^*$ indicates the i-th standardized wavelet coefficient.

2) Selection of wavelet coefficients

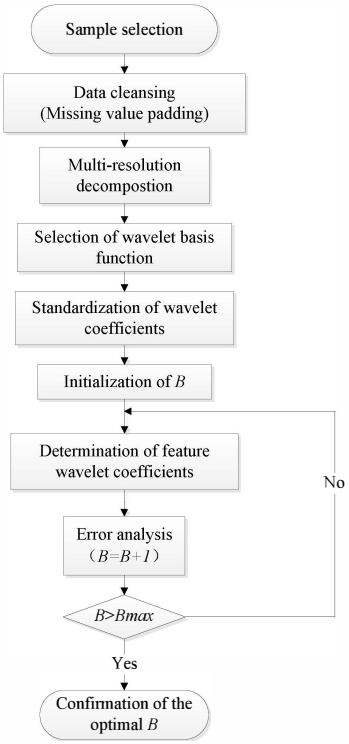


Figure 1. selection of wavelet coefficients

From what studies above, the determination of retained feature wavelet coefficients number B directly affects how much the wavelet coefficients contain the information of the original data, as well as determines the compression ratio of the original data after data characterization. Thus, the determination of the B value is the most critical step in data characterization based on wavelet transform.

In this paper, through the analysis of the error between the data obtained from inverse wavelet transform and the original data, the optimal value of B is determined, so as to achieve transmission simplification, data compression and data desensitization while ensuring the demand of business analysis. Detailed steps are shown in Fig. 1, when different numbers of feature wavelet coefficients are retained, the errors between the corresponding inverse wavelet transform data and the original data are calculated. Through the analysis of the transformation of the error, we can determine the optimal retain number of feature wavelet coefficients B.

III. EXPERIMENTAL ANALYSIS OF ELECTRICITY LOAD DATA

The electricity load data studied in this paper comes from collection of 96 monitoring points per day. Indicator data for single user per day can be regarded as a 96-length vector, and data of n users per day could be formed as a data set with n rows and 96 columns, each line of the data set could indicates a single user. For indicator data of single user multi-days, 96 monitor points per day can be used as 96 variables, monthly data is stored as a 30 (31) × 96 matrix and the annual data stored as a 365 × 96 matrix. In this paper, the monthly (30 days) load data of 1000 users are randomly selected from the power user information acquisition system as samples.

A. Selection of Wavelet Basis Function

Shannon entropy of the wavelet coefficients obtained from different wavelet basis functions is as shown in Fig.2. It can be seen from the figure that, for single-user multi-day electricity load data, the Shannon entropy of the obtained wavelet coefficients is the smallest when d4 (Daubechies 4) is the basis function of wavelet transform; for single-user single-day data, the Shannon entropy of the obtained wavelet coefficients is small when la20 (Least Asymetric 20), bl20 (Best Localized 20) and d4 are the basis functions. Since the d4 wavelet basis function is similar to the structural characteristics of the electricity load data, this paper takes the calculation results of Shannon entropy into account, combines the discrete structure of electricity load data to select the d4 function as the basis function of the decomposition of electricity load data.

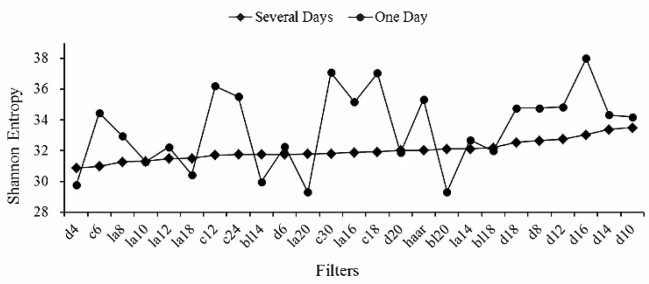
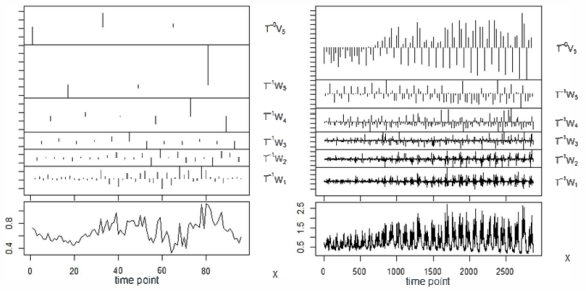


Figure 2. Shannon entropy of wavelet coefficients obtained by using different wavelet basis functions

In this paper, d4 wavelet basis function is selected for single-user single-day data and single-user multi-day data to take wavelet multi-resolution decomposition according to the structural characteristics of electricity load data. As shown in Fig. 3, the higher the decomposition stage is, the more detailed changes of the original data from both the frequency domain and the time domain.



(a) single-user single-day data (b) single-user multi-day data

Figure 3. wavelet multi-resolution decomposition example

B. Error Analysis

The square sum of the error between data obtained from wavelet transform and the original data is taken as the error of data characterization in this paper, modeled as

$$Err_{single} = \sum_{j=1}^{96} (y_j - x_j)^2 \quad (11)$$

$$Err_{multiple} = \sum_{i=1}^{n} \sum_{j=1}^{96} (y_j - x_j)^2 \quad (12)$$

In Equation (11) Err_{single} represents the error for single-user single-day while in Equation (12) $Err_{multiple}$ represents the error for single-user multi-day. n indicates the observation days of single-user multi-day, y_j indicates the sequence data obtained by inverse wavelet transform, and x_j represents the original sequence data.

The error between the inverse transformed data under different B and the original data for 1000 users per day is as shown in Fig. 4 and that for 1000 users per month is shown in Fig. 5.

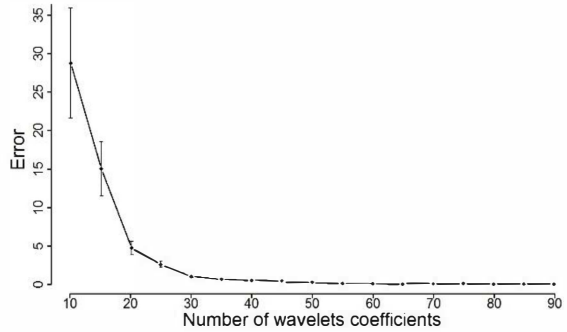


Figure 4. the error for multi-users per day

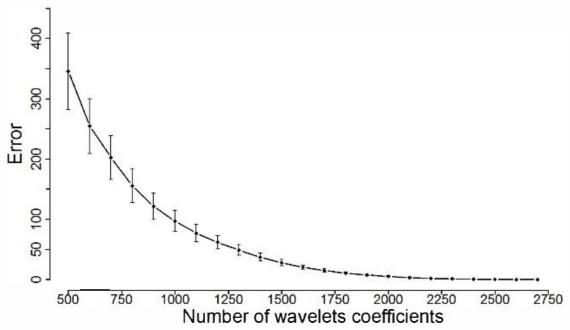
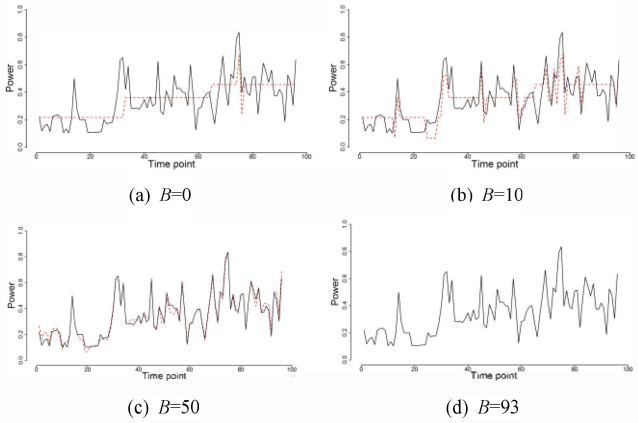


Figure 5. the error for multi-users multi-days

In order to intuitively show the difference between the wavelet transformed data with the original data, Fig. 6 takes a typical user's load curve as an example. When B is settled as 0, 10, 50 and 93, the difference between the inverse wavelet transformed sequence data and the original data is shown in Fig.6.

It is shown in Fig.4 and Fig.6 that the mean value of error decreases obviously with the increasing value of B when B is less than 50, and the mean value of single-user multi-day data does not exceed 0.15 and increasing the number of characteristic wavelet coefficients can not significantly reduce the error between the inverse wavelet transform sequence and the original sequence data when B is not less than 50, which proves that there is a high redundancy in the original power load data and it is necessary to characterize the data.



(a) B=0 (b) B=10

(c) B=50 (d) B=93

Figure 6. the difference between the inverse wavelet transformed sequence data and the original data under different B

IV. CONCLUSIONS

With the rapid growth of global data, information of grid enterprise is also growing rapidly, which causes great trouble in data collection, transformation, storage and application. The massive accumulated data increase grid enterprise operation cost. Therefore, data transformation simplification, data compression and data desensitization have become important issues of grid big data application.

This paper studies the basic principle of wavelet transform and analyzes the application value of multi-resolution decomposition of data characterization. It turns out that the application of wavelet transform in electricity load data for data characterization achieves great effect, including lower data storage space, more simplified transmission and improved analysis efficiency. However, the selections of basis function and wavelet coefficients as characterized coefficients are of vital importance. It will not only affect data compression, but also influences the business analysis result.

The Shannon entropies of the wavelet coefficients obtained from 25 kinds of wavelet basis functions are compared, and d4 function is chosen as the basis function based on the minimum entropy principle. In addition, the inverse wavelet transform is applied to characteristic wavelet coefficients, and the error between the result and the original data is calculated, through the analysis it is proved that the d4 wavelet transform can be used for data characterization to achieve lower data storage, more simplified data transmission and improved analysis efficiency.

Data characterization based on wavelet transform for single-user single-day data and single-user multi-day data achieves great effect on data compression and data desensitization, which makes it possible for big data in power field not only used within power enterprises, but also combine with external data and serve for more fields. Thus the potential value of power data could be reflected.

ACKNOWLEDGMENT

This paper is supported by research project of State Grid Corporation of China SGTYHT/15-JS-191 project.

REFERENCES

[1] Keogh E, Chu S, Hart D, et al. An online algorithm for segmenting time series[C]//Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on. IEEE, 2001: 289-296.

[2] Krishna B V, Baskaran K. Parallel computing for efficient time-frequency feature extraction of power quality disturbances[J]. IET Signal Processing, 2013, 7(4): 312-326.

[3] Zhong Fu Tan, Jinliang Zhang, "A model to forecast electricity price based on multivariate wavelet transform and multivariate time series", Journal of China Electromechanical Engineering, 2010 (1): 103-110.

[4] Park S, Kim S W, Chu W W. Segment-based approach for subsequence searches in sequence databases[C]//Proceedings of the 2001 ACM symposium on Applied computing. ACM, 2001: 248-252. minutes

[5] Hung N Q V, Anh D T. An improvement of PAA for dimensionality reduction in large time series databases[C]//Pacific Rim International Conference on Artificial Intelligence. Springer Berlin Heidelberg, 2008: 698-707. Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," ASME Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[6] Xu X, Kezunovic M. Automated feature extraction from power system transients using wavelet transform[C]//Power System Technology, 2002. Proceedings. Power Con 2002. International Conference on. IEEE, 2002, 4: 1994-1998. Electronic Publication: Digital Object Identifiers (DOIs):

[7] Thirumala K, Umarikar A C, Jain T. A generalized empirical wavelet transform for classification of power quality disturbances[C]//Power System Technology (POWERCON), 2016 IEEE International Conference on. IEEE, 2016: 1-5.

[8] Struzik Z R, Siebes A. Wavelet transform in similarity paradigm[C]//Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer Berlin Heidelberg, 1998: 295-309.

[9] Sun Chong, A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Engineering[D], Huazhong University of Science & Technology,2013

[10] Kumar R, Singh B, Shahani D T, et al. Dual-Tree Complex Wavelet Transform-Based Control Algorithm for Power Quality Improvement in a Distribution System[J]. IEEE Transactions on Industrial Electronics, 2017, 64(1): 764-772.

[11] Ning J, Gao W. Multi-feature extraction for power system disturbances by wavelet transform and Fractal analysis[C]//Power and Energy Society General Meeting, 2010 IEEE. IEEE, 2010: 1-7.

[12] Jagadesh T, Rani R J. A novel speckle noise reduction in biomedical images using PCA and wavelet transform[C]//Wireless Communications, Signal Processing and Networking (WiSPNET), International Conference on. IEEE, 2016: 1335-1340.

[13] Singh S, Jain B, Jain S. Wavelet based real-time power quality monitoring[C]//Electrical and Computer Engineering (CCECE), 2016 IEEE Canadian Conference on. IEEE, 2016: 1-6.

[14] Zheng Ruijiang, Yang Zhenbin, Liu huichao, A medthod of power system harmonic detection based on wavelet transform, Power System Protection and Control, 2012, 40(15) : 35-39

[15] WU Mei, LI Zhong-jian,LIU Xiao-gang,Feature Extraction Based on Wavelet Data Compression Obtained by Wavelet Multi-resolution Analysis, Journal of Projectiles, Rockets, Missles and Guidance, 2006, 26(4): 408-410